DEVELOPMENT OF OBJECT DETECTION TECHNIQUE USING MACHINE LEARNING

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies
(Deemed to be University)

by

Shweta Bali
Registration No:13/Ph. D/013

Under the supervision of

Dr. S.S. Tyagi
Professor, Computer Science and Engineering

FACULTY OF ENGINEERING & TECHNOLOGY

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121004

FACULTY OF ENGINEERING & TECHNOLOGY

Certificate

I hereby declare that the short synopsis "Development of Object Detection Technique using Machine learning" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science and Engineering under Faculty of Engineering and Technology of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2019-2020.

It is a bona fide record of original work carried out under the supervision of Dr. S.S. Tyagi, Professor and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

(Shweta Bali)
13/Ph. D/013

Signature of the Supervisor
Dr. Shyam Sunder Tyagi
Professor, Computer Science & Engineering
Faculty of Engineering & Technology

(Chairman, DRC)

Chairperson, CDP

ABSTRACT

Object detection plays a vital role in different applications including face detection, human behavior analysis, pedestrian detection, image retrieval, photo-tagging of faces, autonomous driving and video surveillance. Object detection is the problem of finding location of objects (object localization) and assigning category (object classification) to each of them in images and videos. Machine learning is a field of Artificial Intelligence (AI) which deals with the study of algorithms and models that computer systems use to perform a certain task. The traditional machine learning techniques for image classification are made up of two sub-processes namely feature extraction process and classification process. To detect an object, the features such as Scale Invariant Feature Transform (SIFT), Haar-like features, Histogram of Oriented gradients (HOG) that are highly representative in nature are used to form a feature vector and during classification process classifiers such as Support Vector Machines (SVM), Deformable part models (DPM), Adaboost are used to assign class labels to different objects. Then the rectangular box is marked around the objects that are classified in different categories. The key challenges faced by these machine learning techniques were redundant bounding boxes, low performance due to large variation in appearance, pose, lightning conditions and different backgrounds. Moreover, it is difficult to develop a generic descriptor for features that describes different kinds of objects. Object detection algorithms have gone through significant improvement since 2012, due to availability of datasets, graphical processing units (GPU's) and pretrained networks the cost and the training time has been reduced drastically. Neural networks with large number of layers have higher capacity to learn complex features and are able to optimize certain tasks. The deep learning models are capable to deal with occlusion, scale transformation so they are the hot spot for research. The research using convolutional neural networks is still evolving but still there are challenges to be handled such as occlusions, variations in pose & viewpoint, object rotation and scale changes, different illumination conditions, small size of the objects, distortion, intraclass variation in the images which are faced by object detection algorithms in different applications. The main aim of research would be the building improved technique to help the applications which require object detection.

Keywords: Convolutional Neural Networks, Object Detection, Machine Learning, Deep Learning

CONTENTS

S. No.	Description	Page No.
1	Introduction	1-3
2	Literature Review	4-11
3	Description of Broad Area	12-16
	3.1 Research Gap	15
	3.2 Research Problem	16
4	Objectives of the Study	17
5	Methodology to be adopted	18
6	Expected Outcome of the Research	19
	References	20-24

1. Introduction

Detecting objects is a key area of research in Image Processing and Computer Vision for detecting location and class of different objects present in images and videos. It has a widespread adoption rate in several applications such as image retrieval, face detection, video surveillance, pedestrian detection, automated tagging of faces. Object detection helps in gaining the understanding of images. Traditional object detection methods employ two phases namely feature extraction and classification. Many representative features have been used in literature such as Scale Invariant Feature Transform (SIFT) that are invariant to scale and rotation by Lowe, D.G. (2004), histogram of gradients (HOG) by Dalal, N., Triggs, B. and Schmid, C. (2006). The feature vectors extracted are later fed to classifiers such as Support vector machines (SVM) and Adaboost which are good choices for identifying the classes of the objects in the class. The main bottleneck in the traditional object detection pipeline is the performance of the classifier is highly dependent on kind of features extracted so generic pipeline is difficult to be generated. The other challenges which made object detection difficult are diversity in appearances and pose, illumination conditions, variation in size and backgrounds. Deep learning has made a great success in the field of object detection due to existence of Convolutional Neural Networks (CNNs).

Object detection techniques have undergone major revisions since 2012. Solutions for object detection problems have been found by conducting challenges such as PASCAL Visual Object Challenge (VOC) and ImageNet Large Scale Visual Recognition Challenge (ILSVRC)**.** Pascal VOC had a tenure from 2005 till 2012. From 2010 onwards, the ILSVRC was an annual event and became the main standard for object detection after Pascal VOC concluded in 2012. ILSVRC challenge uses ImageNet dataset (contains 14 million plus images) which spans across 1000 classes of the objects classification and 200 classes for detection. In 2012, there was huge performance improvement, AlexNet which is trained a deep convolutional neural network (CNN) was used for classification of large dataset of high-resolution images into 1000 different classes. The first breakthrough came in object detection due to network R-CNN (Region-based Convolutional Neural Network) which resulted in an improvement of nearly

30% over the datasets in PASCAL VOC 2012 challenge. The deeper architectures exploited by CNNs extract hierarchical features which help in extracting the useful information. There are several factors such as emergence of large datasets, high performance parallel systems GPUs and advances in the network structures like AlexNet, GoogLeNet, Visual Geometry Group (VGG), and Residual Net (ResNet) that have improved the performance drastically.

Datasets are the foundation stones of deep learning. For the experiments to be carried out successfully large datasets are required. Some of the commonly used open datasets are ImageNet dataset containing more than 14 million images divided into 20,000 categories of the object, PASCAL VOC 2012 dataset which includes 20 classes that contains 11,530 images, COCO (Common Objects in Context) dataset by Microsoft named as MSC COCO dataset that spans over 80 object categories with more than 300,000 images.

The assessment of performance an object detection algorithm is done by comparing the results against the ground truth data. Some of performance metrics that are used to capture different aspects of performance of object detectors are Precision, Recall, mean Average Precision (mAP), Accuracy, Confusion matrix, F1-score, Receiver Operating Curve (ROC) curve, Precision-Recall Curve, Intersection over Union threshold (IoU).In a typical dataset there are a number of classes with non-uniform distribution the accuracy metric is biased and also, we need to find a score with the bounding box so AP (Average Precision) is useful. mAP score is estimated by finding mean value of precision of all recall values. Precision measures the fraction of true object detections to total number of predicted objects by the classifier. Recall measures the fraction of true object detections to the total number of objects in the data set. To assess the task of object localization the model is evaluated on the Intersection over Union threshold (IoU).

2. Literature Review

Study in the area of object detection has attained a state where the results are used in real-time applications. There are several challenges that object detection system has to cope with such as effects of cluttered background, variation in illumination, varied positions and scales in which objects appear in different images. To handle these complexities the researchers have designed powerful descriptors, ensemble-classifiers, and different machine learning architectures. Early 90s used appearance-based techniques, mid 90's there came sliding window-based approaches, late 90's feature-based approaches, early 2000s parts and shape-based techniques,2003 onwards bag of features (BOG) based techniques,2012 onwards deep learning techniques.

Papageorgiou, C. and Poggio, T. (2000) used example-based techniques to detect humans as full body object. These systems were successful in detecting front, side and rear views but failed in partial occlusion.

Viola, P. and Jones, M. (2001) for face detection used integral images which speeded up the convolution process. The authors used Adaboost for selecting small number of important features. The authors also reduced the computation time by constructing cascade, cascading of classifiers that lead to improved performance. They also provided experimental results on face dataset with variation in illumination, pose, scale and camera.

Dalal, N., Triggs, B. (2005) proposed a shape-based feature descriptor named Histogram of oriented gradients (HOG), which reveals both appearance as well as shape of object. These are feature invariant and are computed on dense grid of cells. The descriptor is capable of detecting a variety of the objects but primarily they are used for pedestrian detection. They also discussed that the parameters such as fine-scaled gradients, local contrast normalization, orientation binning, coarse spatial binning are all significant for good performance. They also introduced a challenging dataset for detecting the pedestrians.

Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008) proposed deformable part-based model (DPM) for solving generic object detection problem. The model achieved improvement

in the performance on PASCAL person detection dataset. Discriminative procedure is used for training the model. The technique works by learning from the weakly labelled data. These models are good at detecting the partially occluded objects. The authors suggested further scope to learn deeper hierarchies i.e., parts within parts ,3D pose, detection of multiple classes using the same vocabulary of the parts. They further recommended to use A* search algorithm to fetch the latent variables during detection process.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) proposed GPU efficient implementation of 2D convolution process, new features were added. They also found that different data augmentation and dropout techniques reduce overfitting and thus generalizing the model. The network consists of five convolutional layers along with three fully connected layers. They also proposed GPU efficient implementation of 2D convolution operation. As there are 60 million parameters so it is difficult for the network to learn so may parameters without overfitting. Data augmentation and dropout techniques were used to reduce overfitting on the images and thus generalizing the model. The network architecture consists of five convolutional layers and three fully connected layers. The experiments brought great amount of reduction in Top 5 classification accuracy and localization error. The study suggested use of larger and deeper convolutional nets to the videos to extract more information. They also suggested using unsupervised pre-training to simplify the experiments.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) came up with a technique called Region-based CNN(R-CNN) which used region proposals to localize and segment the objects. They used the algorithm based on selective search for extracting 2000 regions present in images. These regions were fed to convolutional neural network and feature vectors extracted from these regions and fed to SVM classifier to predict class of the object. They concluded that there is sharing of CNN parameters across different classes. Also, the dimensions of the feature vector have reduced drastically. They also showed that pretraining the network with supervision is highly effective and then the network is fined tuned for the target class for small data. R-CNN achieves mean Average Precision (mAP) on PASCAL VOC07 58.5%. The limitations of the study are high feature dimensions and slow speed due to large no. of the region proposals. After that, the research in the field of object detection gained momentum.

Other factors for the success include pre-trained models, large sized benchmarked datasets such as ImageNet**.**

He, K., Zhang, X., Ren, S. and Sun, J. (2015) proposed a technique which uses pooling strategy which eliminates the need of fixed size input image. Spatial pyramid pooling network (SPP-net) is developed by building the SPP layer above the last convolutional layer. The method computes convolution feature map for the full image and uses it to classify each object proposal. As compared to R-CNN there is improvement in the speed as the convolutional layer has to run once for the full image. The proposed technique handled different scales, sizes, and aspect ratios. The study also revealed that the network achieved remarkable performance in classification and reduction task.

Zeiler, M.D. and Fergus, R. (2014) introduced a visualization technique that was used to find problems to obtain better results provides detail of the intermediate layers and the classifier. The visualization technique used multi-layered deconvolutional network (deconvnet). ImageNet trained model performed well on datasets such as Caltech-101, Caltech-256 whereas for trained models built using smaller datasets such as PASCAL, these models generalize less but the performance can be enhanced by using various types of loss functions that are used to detect multiple objects per image.

Simonyan, K. and Zisserman, A. (2014) investigated the effect of the depth of the convolutional neural networks on the accuracy of the recognition task. They proposed a network architecture which used 3*3 filter and achieved significant improvement by using the depth of 16-19 layers. The architecture used smaller receptive fields of size 3*3 throughout the whole network with stride of 1. The implementation allows performing training as well as evaluation on multiple GPUs in which batches of training images are split and processed on parallel GPUs. The experiments are performed on ILSVRC-2012 dataset and top-1 and top-5 errors are computed for evaluating the performance. The authors found that increasing the depth from 11 to 19 layers causes decrease in the classification error. The VGG-Net based network beats the previous networks.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) presented a new architecture named Inception which improvised the resource utilization in the network. They work based on Hebbian principle that is neuron. that fire together, wire together. The main consideration that is considered is that the performance of the deep neural networks can be improved by increasing both depth as well as width of the network. Due to increase in the network size, number of parameters also rise that causes the network to overfit, usage of computing resources is also increased drastically. The authors suggested that these issues can be solved by using sparse connected architectures. But the challenges again were that the existing infrastructures are inefficient for sparse architectures and the libraries use the dense matrix multiplication. The authors presented an inception architecture called GoogLeNet which implements the sparse structure which can be run on individual devices. The total number of the layers are 100 out of which 22 layers contain parameters. Average pooling has been used to improve the results. The network is trained using asynchronous stochastic gradient algorithm. It achieves top-5 error rate 6.67% on the dataset. The study provided the evidence that sparse structures can be approximated using dense building blocks and helps in reaching improved network performance.

Girshick, R. (2015) presented an algorithm named Fast-RCNN which has achieved high detection quality measured by calculating mean average precision (MAP score) as compared to previous approaches. The main overhead in RCCN was the process of running different region proposals through the convolutional base for each region proposal. This technique removed this overhead in which the input image is fed to Convolutional neural network as an alternative to region proposals to create a convolutional feature map. From the convolutional feature map, the region proposals were extracted using selective search algorithm. Softmax layer was used to predict category of proposed regions and also to calculate offset values of bounding box. The main contribution of this architecture is sharing features of convolutional network. It also introduces a single pipeline for classification and regression which reduces the training time of the network exponentially.

Ren, S., He, K., Girshick, R. and Sun, J. (2015) introduced a replaced selective search algorithm by Region Proposal Network (RPN) to find proposals of regions which are further

reformed using a Region of Interest (ROI) pooling layer. RPN uses anchors to predict the regions of multiple scales and aspect ratio across the image. The layer is then used for classification of the image within the proposed region and prediction of offset values for different bounding boxes. Small objects are difficult to detect as they have low resolution and most of the features are lost due to convolution and pooling operations of CNN. The authors concluded that that the network is achieves detection at a faster rate.

Razakarivony, S. and Jurie, F. (2016) introduced a new database VEDAI (Vehicle Detection in Aerial Imagery) of more than 1200 small aerial images with a total of more than 3700 annotated targets that exhibit different variations such as shadow, occlusions, and lightning changes. The experiments are performed on different features namely HOG (Dalal, N., Triggs, B. ,2005), LBP (Ojala, T., Pietikäinen, M. and Mäenpää, T., 2002), LTP (Tan, X. and Triggs, W. ,2010), DPM (Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2009) and classification is performed by Support Vector Machine. They also performance on different large sized images as well as color images and found that the performance is better on colored images and resolution is major factor for improvement of the performance also the performance of the classifier can be achieved by adding hard negative examples. They concluded that further research can be done to develop new algorithms on the developed dataset.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) introduced a neural network to achieve scale invariance. These networks are known as Feature Pyramid Networks (FPN) which generate multiple feature maps to represent an image at different scales. Later Faster-RCNN is run on each scale and multiscale testing is performed. The authors inferred that pyramid representations are required to address multiscale problems.

Chen, G., Choi, W., Yu, X., Han, T. and Chandraker, M. (2017) proposed the new framework with knowledge distillation which improved the classification results and hint learning. Different datasets on that are evaluated include g PASCAL, KITTI, ILSVRC and MS-COCO. The authors observed that applying distillation techniques to multi-class object detection is challenging because results deteriorate due to compression, class imbalance poses

another challenge to the problem, The authors claimed that distillation injects structural information which improves the generalizability of the model and hint learning helps in solving under fitting problem .The study revealed the underfitting occurs in CNN and be further explored to improve understanding of learning mechanisms for object detection .

Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. and Ouyang, W. (2017) proposed a new deep learning framework for object detection in videos. The videos contain the temporal information as well as contextual information that needs to be addressed. They performed the analysis on the ImageNet VID challenge. This can be further explored.

Lin, X., Zhao, C. and Pan, W. (2017) used a scheme called ABC-Net that used binary neural networks (BNNs) which are networks with binary weights and activations that can take two binary values (-1 or +1). They are advantageous as they reduce memory usage and model size decreased by 32, reduced power consumption and replacement of arithmetic operations by speeded up bitwise operations. The authors concluded that this architecture is useful for mobile applications and real-time embedding systems. ABC-Net architecture can be implemented on FPGA or any other deep processor.

Cao, J., Pang, Y. and Li, X. (2017) researched on a framework called Multilayer channel features which combines HOG and LUV features in each CNN layer and then classified using multistage cascaded Adaboost classifier. They are based on the notion that each layer in CNN contains different features which are used by learning algorithm to efficiently classify the objects and the backgrounds. The experiments are evaluated on Caltech, ETH, INRIA ,TUD-Brussels and KITTI data set. To further enhance the performance the author's suggested that the overlapped windows with low scores are eliminated after the first stage. It also reduces the computation cost of operations of CNN. The study concluded that work needs to explored on very thin networks to achieve remarkable results.

Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z. and Zhou, H. (2017) examined the use of local as well as contextual visual cues to

improve object detection. The authors proposed a network known as GBD-Net which used gates for passing messages among features present in different regions with different resolutions using convolution in different layers. The experiments were performed on three sets namely ImageNet, Microsoft COCO, and Pascal VOC2007.The main concept being used is that the surround regions provide the contextual information about the objects in the vicinity as well as the background.

Hu, Q., Wang, P., Shen, C., van den Hengel, A. and Porikli, F. (2017) researched that the convolutional neural networks performed poorly for detecting pedestrians. They stated that deep convolutional neural networks (DCNN) with number of features (optical flow and ACF features) and mechanisms (pixel labelling) help in performing better. They used convolutional feature maps (CFMs) of DCCN as features extracted to train ensemble of decision tree classifier. The results were performed on Caltech pedestrian dataset with log-average miss rate of 8.9%. The authors suggested that ROI pooling can further increase the speed and hence the performance.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) discovered that the main cause of accuracy drop in two-stage approach for object detection is due to extreme class imbalance between the classes. They solved the problem by redesigning the standard cross entropy loss called focal loss that focusses on learning on hard samples. The detector is called RetinaNet which is a one-stage detector that achieved better results (mAP score) on the COCO dataset than the two stage detection algorithms by scaling the entropy loss such that the scaling factor reaches 0 as the confidence in the correct class increases.

Sun, X., Wu, P. and Hoi, S.C. (2018) investigated a new technique for detecting faces by extending the Faster R-CNN algorithm and proposed the practices to perform certain tasks such as hard negative mining, multi-scale training, feature concatenation, model pretraining, calibration of key parameters. WIDER FACE dataset is used for training and validation VGG16 which has been trained on ImageNet was selected as the model. The results are evaluated on FDDB dataset (Jain, V. and Learned-Miller, E. (2010)).

Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G. and Wu, J. (2018) predicted the problem to detect small objects using the Single Shot Multibox Detector (SSD) which is based on VGG16 base network. It uses the pyramid hierarchy of the features to predict the objects at different scales and use the shallower layers to predict the smaller objects Two types of feature fusion methods namely concatenation and element sum module are used. The model capable to detect small objects at a faster speed and with comparable accuracy. The contextual information is useful for improving the results but sometimes adds unnecessary noise that needs to be dealt. The authors concluded that shallower layers predict smaller objects and lack the semantic information which if given to the convolutional network will improve the detection of small objects.

Sommer, L., Schuchert, T. and Beyerer, J. (2018) performed research by taking aerial images for object detection into consideration. They investigated two frameworks namely Fast R-CNN and Faster R-CNN which have achieved remarkable performance in object detection. These networks were able to perform better than the existing networks in detecting small instances of vehicles. They met different challenges such as variation in size, number of objects per image, varying backgrounds. The authors claimed that their network using Faster R-CNN performs better than the existing works on the two datasets namely DLR 3K Munich Vehicle Aerial Image Dataset (Liu, K. and Mattyus, G., 2015) and Vehicle Detection in Aerial Imagery (VEDAI) dataset (Razakarivony, S. and Jurie, F., 2016). The authors proposed that the detection frameworks can be explored for rotated bounding boxes.

Hu, G.X., Yang, Z., Hu, L., Huang, L. and Han, J.M. (2018) proposed multiscale deep convolutional neural network based on Faster R-CNN to detect small objects. The model is composed of four layers namely feature extraction layer, feature combination layer, RPN (Region Proposal Network) layer and the classification and regression layer. The model attained remarkable progress in detecting the small objects.

3. Description of the Broad Area

Object detection is the technique for estimating the class as well as location of the objects present in the image. The idea behind object detection is building models and techniques for understanding images and videos. Machine learning help computers to learn automatically from data without human interaction and improve from experience. Machine learning techniques can be broadly divided into supervised and unsupervised algorithms. Supervised techniques use labeled data from the past to predict the outcomes for future events whereas the unsupervised techniques learn from the unlabeled data. Supervised techniques are divided into classification and regression tasks. In classification, the algorithm forecasts class of new data from training dataset. In regression the algorithm predicts the continuous output. The traditional machine learning algorithm first performs feature extraction from the data provided during training which is later given to the classifier for predicting class of the objects. Since the training data cannot incorporate each and every feasible input so the learning algorithm must be trained to predict the output for unseen instances. The output of the algorithm is evaluated by using the loss function and the aim of training is to minimize the loss. Deep learning is a subgroup of machine learning which utilizes Artificial Neural Networks (ANN)s to accomplish machine learning tasks. The deep learning techniques utilize convolutional neural networks as the backbone for developing the powerful models. A lot of research is being undertaken on improving the deep learning architectures. There are few basic steps as shown in Figure 1. that each algorithm performs to achieve the object detection. They are stated as :-

1. **Creation of dataset**: A dataset will be created for the required application. Perform processing operations on the dataset. Some of the algorithms use existing benchmarked datasets whereas some design new datasets according to the need.

2. **Preprocessing:** These include techniques for handling the missing values, standardizing the data, encoding categorial values if desired, splitting the data into training and testing dataset which are required to transform data in the dataset /images into the format desired by the underlying model or architecture.

3. **Feature Extraction**: This step involves application of algorithms to extract features from images based on data as well as the problem at hand. As the number of the features extracted by different algorithms are large in number some dimensionality reduction techniques need to be applied to those features. In literature many techniques are available out of that principal component analysis (PCA) is one that is used extensively. In deep learning architectures the features are extracted automatically using convolutional neural networks (CNNs) such as VGG, GoogLeNet.

4. **Classification**: This involves the algorithms that are used to predict class/labels of objects present in images or videos. There are several classification techniques such as decision tree, Naïve Bayes classifier, (K-Nearest Neighbors) KNN, Multilayer perceptron (MLP), Convolutional Neural Networks (CNNs) that can be used.

5. **Bounding-Box Computation**: In object detection bounding box is a rectangular box that describes position of the target object. Various techniques such as regression are available in literature to perform this task.

6. **Evaluating Performance & Optimization:** Several metrics such as Intersection over union (IOU), Mean average precision (mAP), Precision, Recall, F1-score, ROC Curves to name a few are used to evaluate the performance and some parameters are also tuned to optimize the results.

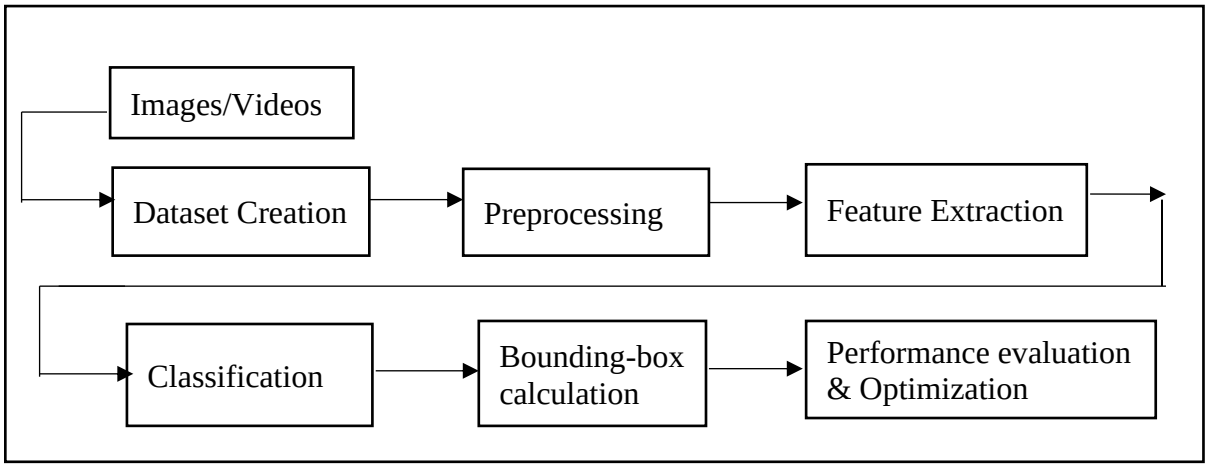

Figure 1: General Steps for object detection

There are several challenges faced by the detection algorithms. Some of them are:

1. Occlusion: It refers to the problem of one object hidden by the other object which causes the feature extraction to be difficult.
2. Different Viewpoints: In different viewpoints the shape of the object changes and the features of the objects also changes which causes difficult to detect the objects.
3. Variation in size: Different objects have different sizes as the same object can be far or near to camera.
4. Blurred objects: Due to camera motion the images of some of the objects get blurred which causes the detector to fail.

Some of the benchmarked datasets used in Object Detection, Classification and Recognition are listed in Figure 2. that are extensively used in literature.

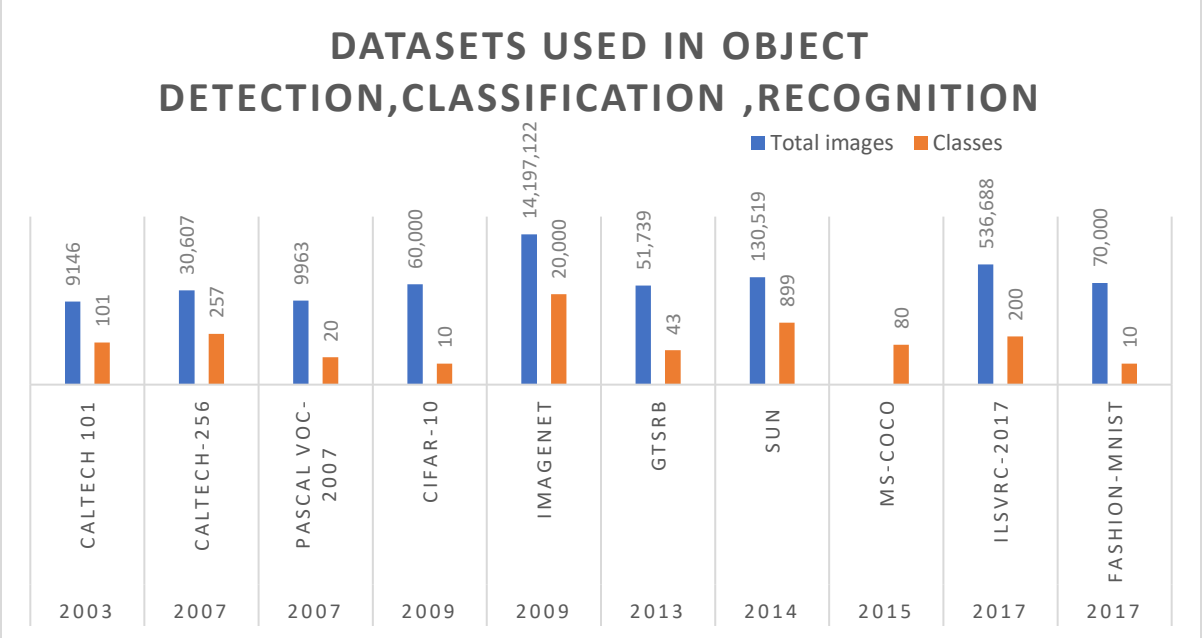

Figure 2: Datasets used in object detection, classification & recognition

3.1 Research Gaps

Each research problem arises due to some issues that are not addressed or are challenging and needs to be improved. In the area of object detection also there are number of such issues. Some of the gaps found are stated as: -

1. Machine performance is still low especially for detecting some small objects so the research can be done in this direction.
2. There is a huge need for developing detection algorithm that run smoothly on mobile devices.
3. Deep learning-based techniques are sophisticated, and the need is to reduce human interference in designing the detection model.
4. Object detection for some real-world applications, remains a challenge. Generative Adversarial Networks (GAN) can be one area for object detection in the future.
5. Detecting small objects in large scene is potential application of this research direction includes counting the population of wild animals with remote sensing images and detecting the state of some important military targets.
6. Real-time object detection in videos is of great importance for video surveillance and autonomous driving. Traditional object detection algorithms are usually designed using images which ignores correlations between videos frames. Improving detection by exploring the spatial and temporal correlation is an important research direction.
7. Object detection with multiple sources of data, e.g., RGB-D image, 3d point cloud is one of the future areas of research for autonomous driving and drone applications.

3.2 Problem Statement

This research aims to develop a technique for detecting the objects under different challenges like variations of pose, scale, location, illumination changes, intra-class variation, variety of scales, developing mobile based application, detecting small objects to improve the performance under various challenging situations.

4. Objectives of the Study

To develop a technique for object detection in images using machine learning techniques to assess its performance in terms of various parameters. The research objectives are outlined as:

1. To generate the dataset for the required application.
2. To review the technologies and methodologies for object detection in images and apply existing algorithms on the developed dataset.
3. To develop new technique for efficiently detecting the objects.
4. To compute the performance measure of the developed technique.
5. To compare the proposed with existing techniques to investigate the performance on existing and new dataset.
6. To optimize the technique using different tweaks to improve the performance.

5. Methodology to be adopted

The major steps that are involved to achieve the above research objectives are summarized and detailed below: -

1. A detailed literature review will be done on the existing algorithms, models built, and the technologies used for detecting objects in images.
2. The study done in literature review will shortlist the research needs. The gaps found will justify the development of the detection technique for the application. Then the need of research and research objectives will be defined.
3. Research objectives will be defined along with the problem statement.
4. After the need for research has been identified & research objectives have been defined, existing object detection algorithms will be implemented in Python/MATLAB and their performance will be evaluated in terms different performance measures on new and existing datasets. This will help in getting into the insights of the research requirement.
5. A new object detection technique will be developed for desired application and the technique will be evaluated on the existing as well as new dataset in order to check the performance under different challenges.
6. Further optimization will be applied to developed technique to improve the performance.

6. Expected Outcome of Research

Research has been extensively done around detecting objects. But still there are certain issues such as occlusions, different illumination conditions, variations of pose, scale, location, illumination changes, intra-class variation, variety of scales, developing mobile based application, detecting small objects. The outcome of research is to develop a technique that is computationally efficient in terms of different performance measures in the real-time environment.

References

Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G. and Wu, J., 2018, April. Feature-fused SSD: Fast detection for small objects. In *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)* (Vol. 10615, p. 106151E). International Society for Optics and Photonics.

Chen, G., Choi, W., Yu, X., Han, T. and Chandraker, M., 2017, December. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 742-751).

Dalal, N., Triggs, B. and Schmid, C., 2006, May. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision* (pp. 428-441). Springer, Berlin, Heidelberg.

Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), pp.303-338.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), pp.1627-1645.

Felzenszwalb, P., McAllester, D. and Ramanan, D., 2008, June. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Girshick, R., 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), pp.1904-1916.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hu, Q., Wang, P., Shen, C., van den Hengel, A. and Porikli, F., 2017. Pushing the limits of deep CNNs for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(6), pp.1358-1368.

Hu, G.X., Yang, Z., Hu, L., Huang, L. and Han, J.M., 2018. Small object detection with multiscale features. *International Journal of Digital Multimedia Broadcasting*, *2018*.

Jain, V. and Learned-Miller, E., 2010. *Fddb: A benchmark for face detection in unconstrained settings* (Vol. 2, No. 4, p. 5). UMass Amherst technical report.

Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. and Ouyang, W., 2017. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), pp.2896-2907.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, pp.1097-1105.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.

Lin, X., Zhao, C. and Pan, W., 2017. Towards accurate binary convolutional neural network. *arXiv preprint arXiv:1711.11294*.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), pp.91-110.

Ojala, T., Pietikainen, M. and Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), pp.971-987.

Papageorgiou, C. and Poggio, T., 2000. A trainable system for object detection. *International journal of computer vision*, *38*(1), pp.15-33.

Razakarivony, S. and Jurie, F., 2016. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, *34*, pp.187-203.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Ren, Y., Zhu, C. and Xiao, S., 2018. Small object detection in optical remote sensing images via modified faster R-CNN. *Applied Sciences*, *8*(5), p.813.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), pp.211-252.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sommer, L., Schuchert, T. and Beyerer, J., 2018. Comprehensive analysis of deep learning-based vehicle detection in aerial images. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(9), pp.2733-2747.

Sun, X., Wu, P. and Hoi, S.C., 2018. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*, *299*, pp.42-50.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Tan, X. and Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, *19*(6), pp.1635-1650.

Viola, P. and Jones, M., 2001, December. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.

Viola, P. and Jones, M.J., 2004. Robust real-time face detection. *International journal of computer vision*, *57*(2), pp.137-154.

Zeiler, M.D. and Fergus, R., 2014, September. Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham.

Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z. and Zhou, H., 2017. Crafting gbd-net for object detection. *IEEE transactions on pattern analysis and machine intelligence*, *40*(9), pp.2109-2123.

Shweta Bali Short Synopsis Sep 28 2019

ORIGINALITY REPORT

Excluding Matches less than 14 wo...

7%	6%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Shweta Bali
3/10/2019

PRIMARY SOURCES

03/X/19

Dr. Vikram Sharma
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research & Studies
Faridabad

1	export.arxiv.org	2%
2	acervodigital.ufpr.br Internet Source	2%
3	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
4	"2018 IndexIEEE Transactions on Circuits and Systems for Video TechnologyVol. 28", IEEE Transactions on Circuits and Systems for Video Technology, 2018 Publication	<1%
5	www.ee.oulu.fi Internet Source	<1%
6	Submitted to University of Northumbria at Newcastle Student Paper	<1%
7	doras.dcu.ie Internet Source	<1%

Scanned with CamScanner

8	conradsanderson.id.au Internet Source	<1%
9	Sebastien Razakarivony, Frederic Jurie. "Vehicle detection in aerial imagery : A small target detection benchmark", Journal of Visual Communication and Image Representation, 2016 Publication	<1%
10	multidrone.eu Internet Source	<1%
11	www.groundai.com Internet Source	<1%
12	Submitted to 54339 Student Paper	<1%
13	"Image Analysis and Recognition", Springer Science and Business Media LLC, 2019 Publication	<1%

Exclude quotes On

Exclude bibliography On

Exclude matches < 14 words

Scanned with CamScanner